An empirical study of the Algerian dialect of Social network

Karima Abidi, Kamel Smaïli

SMarT Group, LORIA, F-54600, France
{karima.abidi, kamel.smaili}@loria.fr

Abstract

In this paper, we present analysis on the use of Algerian dialect in Youtube. To do so, we harvested a corpus of 17M of words. This latter was exploited to extract a comparable Algerian corpus, named CALYOU by aligning pairs of sentences written in Latin and Arabic. This one was built by using a multilingual word embeddings approach. Several experiments have been conducted to fix the parameters of the Continuous Bag of Words approach that will be discussed in this article. The method we proposed achieved a performance of 41% in terms of Recall. In the following, we present several figures on the collected data that led to several unexpected results. In fact, 51% of the vocabulary words are written in Latin script and 82% of the total comments are subject to the phenomenon of code-switching.
Index Terms: Algerian dialect, Code-switching, comparable corpora, Word embedding.

1. Introduction

Modern Standard Arabic (MSA) is the official language shared by the entire Arab world that is a simplified form of the old Arabic (Classical Arabic), this last one is found only in the religious texts. Beside of MSA, there is another form of Arabic widely used, but it is generally dedicated to the daily communications, named Arabic dialect or *Darija* in Maghrebi countries. Nowadays, with the advent of social networks, the Arabic dialect is widely used, because the Maghrebi people prefer posting their messages in their local colloquial instead of MSA. Arabic Dialect is henceforth written, it arises several new NLP issues. In fact, this form of Arabic undergone a great morpho-syntactic modifications by relaxing several grammatical constraints of MSA. Furthermore, each Arabic region has its own dialect, which leads to different linguistic variations. There is a large number of Arabic dialects: Arabian Peninsula, Levantine, Mesopotamian, Egyptian, and Maghrebi.

In this work, we are interested by the Algerian dialect that has several specificities, among them, the use of a lot of French words and since recently, English words due to the new immigration of Algerian people to English-speaking countries.

In a previous work [1], we addressed the difficult issue of creating not only a dialectal corpus, but a comparable corpus. In fact, parallel or comparable corpora are an essential material for several NLP applications, such as machine translation, building bilingual dictionaries, and so on. In [1] we proposed to build automatically a Comparable spoken ALgerian corpus extracted from YOUtube (CALYOU). The method proposed is based on the concept of learning multilingual word embeddings (Word2Vec).

In this paper, the objective is to make a deep analysis concerning the collected corpus. This will help to understand how the Algerian people write in social networks and to analyze the important phenomenon of code-switching. In fact, people switch from the Algerian dialect to MSA, French and sometimes to English.

The rest of the paper is organized as follows: Section 2 concerns the related work, while Section 3 we describes the collected corpus and we conduct a deep analysis on these data. Section 4 discusses the automatic multilingual word embedding method used to develop CALYOU, we present also in this section some experimentations to set up Word2Vec parameters. In Section 5 we present some figures concerning CALYOU and then we conclude.

2. Related Works

The NLP community started few years ago to pay attention to Arabic dialects processing. However, at the beginning the majority of these works has been limited only to some dialects and mainly to develop NLP tools rather than resources. In this section, we will present a global overview of research related to Arabic dialect processing with a focus on Algerian dialect corpora. Building resources such as corpora or lexicon is a time consuming process, especially for complex or vernacular languages. For Arabic dialect several researches handle the issue of developing resources. In [2], authors created parallel corpora by using crowdsourcing approach to translate sentences from Egyptian and Levantine into English. A multi-dialect Arabic (MSA and dialects from Egypt, Arabic Peninsula and Levantine) speech parallel corpus has been proposed in [3]. This kind of resources is very rare, since it is expensive and time consuming. In fact, 32 speech hours have been recorded that corresponds to 1291 recordings for MSA and 1069 for dialects. Mubarak and Darwish in [4] used Twitter to collect an Arabic multi-dialect corpus for Saudi Arabian, Egyptian, Algerian, Iraqi, Lebanese and Syrian dialect. In [5], the authors presented a multi-dialect Arabic parallel corpus (2000 sentences): Egyptian, Tunisian, Jordanian, Palestinian, Syrian, MSA and English. At the best of our knowledge, there is no work consisting of aligning parallel corpora for Algerian dialect except PADIC (Parallel Arabic DIalect Corpus) [6][1], which covers beside MSA, six Arabic dialects : Annaba (Algerian), Algiers (Algerian), Tunisian, Moroccan, Syrian and Palestinian. This interesting resource has been used to launch the first ma-

[1]PADIC can be downloaded from `http://smart.loria.fr/pmwiki/pmwiki.php/PmWiki/Corpora`

chine translation for Algerian dialect [7],[8] and could be used also in several other applications.
In a previous work [1], we developed automatically a comparable spoken Algerian corpus. The interest of this resource is that, it is trained automatically on data extracted from social networks, in the opposite to what has been done in PADIC or in other resources. In this work, CALYOU has been updated, it includes: Algerian dialect, MSA, French and English.

3. Collected corpus

To build an Algerian dialect corpus, we harvested comments posted by Algerians corresponding to Youtube videos, by using the Google's API [2]. To ensure that the comments collected mainly concern topics posted by Algerians, we chose few keywords to form queries to retrieve videos concerning national news, Algerian celebrity, local football, etc. Table 3 shows some figures before and after preprocessing the collected data, where $|C|$ is the number of comments, $|W|$ is the number of words and $|V|$ is the vocabulary size. We can mention that after the cleaning

	Raw corpus	Cleaned Corpus
$\|C\|$	1.3M	1.1M
$\|W\|$	20M	17.7M
$\|V\|$	1.3M	0.99M

Table 1: The collected YouTube Algerian Dialect Corpus.

process, the corpus has been reduced by around 15% and the vocabulary by around 24%.

3.1. Investigating Algerian Youtube Corpus

In the following, we will present a study concerning the collected corpus. To our knowledge, there is not a recent study about the Algerian dialect used in social networks. The objective of this study is to have an idea about the characteristics of the Algerian dialect. In the first and second lines of Table 2, we present

	Youtube	Percentage
$\|LS\|$	557K	47%
$\|AS\|$	623K	53%
$\|FR\|$	15457	1%
$\|AR\|$	88982	8%

Table 2: Figures on Youtube Algerian comments

the number of comments written respectively in Latin Script (LS) and Arabic Script (AS). The table shows that 47% of the comments are written in Latin Script. This high rate could be explained by the fact that people are influenced by the French culture and also, in Algeria the mobile phone keyboards are by default in French, which makes writing in Latin script easier, even if it is possible to configure the mobile phone in order to have an Arabic keyboard. We remind that, comments in LS correspond to either Arabizi, French or English. Similarly, comments in AS correspond to MSA or dialect. The proportion of French comments in the total corpus and in the subset of the corpus where the comments are written in Latin script are 1% and 2.7% respectively. A comment in LS is considered as French, if each of its words is French found in a dictionary of 6 millions of words [9]. Similarly, 14.2% (which corresponds to 8% of the total number of comments) of the Arabic posts correspond to MSA comments. Such as for French, we used a MSA dictionary of 9 millions of entries [10]. Table 3 gives the length of comments,

	AS	LS
Min	2	2
Max	5211	4046
$Mean$	15	13

Table 3: Some statistics on the extracted comments

for both Arabic and Latin script. We can remark that comments written in AS or LS are in average almost similar in terms of length.
The pie chart of Figure 1 gives details about the distribution of the vocabulary's words. The Modern Standard Arabic represents 21% of the vocabulary, but this result is biased, since a dialectal word may exist in MSA, but it may have a different meaning. For instance, the MSA word شابة means *young*, but in Algerian dialect it means *beautiful*. That is why this rate is not accurate, it is, in fact, difficult to estimate it since we do not have an Algerian dialect lexicon.
Words which are not in a MSA dictionary are considered such as dialectal words, they represent 74% of the whole dictionary. Among them, 46% are written in Latin script. The harvested corpus is composed by 0.99M of distinct words, 51% of them are written in Latin script. This illustrates the impotant use of Latin script in the Algerian dialect used in Youtube. One can remark through this experiment that people prefer writing in Latin script and in addition they do not pay special attention to grammar. Furthermore, for uneducated people, they write a word as they want, or at the best, such as it is pronounced. This probably explains the high number of words of the vocabulary written in Latin script. This diversity of words is illustrated by the

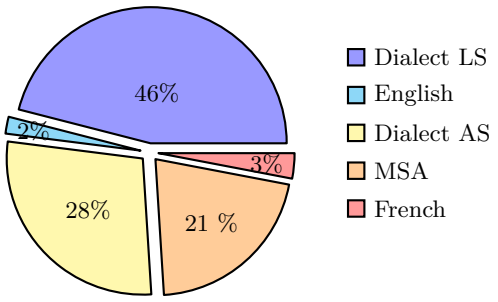

Figure 1: Vocabulary

example (يرحمك) that has been written in 66 different ways in our corpus (see Table 4).

[2]Available at: https://developers.google.com/YouTube

يرحمك → yr7mk yr7mak yrhamak yarhamek yarhemak yarhamk yr7mek yere7mek yarhamak yarhemek yrhmk yar7mak yarhmak yarhmek yar7mik yarahmak yar7mek yarahmk yerhemek yarahmek yerehmek yerhamek yer7mak yare7mek yerhamak yer7mek yerehemek yarhmeke rahimaka yrahmek yrahmak irahmak irhmak irahmek yra7mk irhmk yrehmak yera7mak yerehmk yera7mek yrehmek yara7mak yarehmek yara7mek yerahmeke yarehmak yarhmk yerhmk yarhmeek yra7mak ir7mak yra7mek yarhamoka yrehmk yar7mk yrahmk ira7mak irehmek yerhmek yerahmk yrhmek yerahmek yerhmak yerahmak yrhmak yarahemek

Table 4: Different ways to write in Latin Script the word يرحمك

The pie chart of Figure 2 illustrates an important result concerning the influence of code-switching in the Algerian dialect. In fact, 82% of the comments are a mixture of several languages (MSA, dialect, French and sometimes English). The posts, which are entirely in dialect constitutes only 9% of the total corpus. This proves that the processing of the Algerian dialect necessitates particular NLP tools. In comparison to Hindi, for which the issue of code-switching is also important, the rate of Hindi language, in a corpus of 30 minutes [11] is high (67.7%), while in our corpus the percentage of Algerian dialect is only 39%. The phenomenon of code-switching is crucial, since if we add up all what it has been written in Arabic (Dialect (Arabizi or not) and MSA) in this corpus, only 15.5% of the comments are entirely written in Arabic. All the others are mixture of several codes.

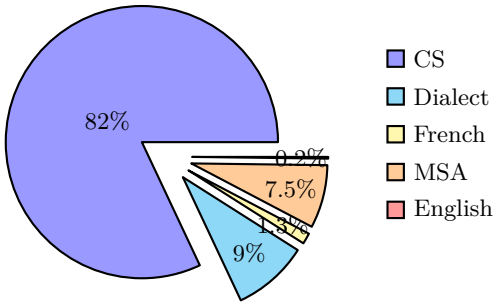

Figure 2: Code-switching distribution

4. Multilingual word embeddings to build CALYOU

In a previous work [1], we addressed the difficult issue of matching comments from YouTube for a vernacular language (Algerian dialect) for which no writing rules do exist. This leads to more difficulties in the processing of the corresponding texts. We recall in this section the main idea of the method proposed in [1]. The comparability of comments cannot be addressed, only by looking for a word into two different comments. In fact, a word, as explained in this paper, may have several ways of writing. That is why, in this method we decided to find for each word, the corresponding ones.That means all the entries, which are correlated to this word and those which are similar but are written differently (see the example of Table 4) constitutes a set, in which the algorithm of comparability looks for the matching. The proposed approach is based on the concept of learning multilingual word embeddings (Word2Vec). The objective is to build a lexicon that contains, for each word its correlated words. To learn a Correlated Words Lexicon (CWL), for each word (w_s) of the corpus, where s is the Arabic or the Latin script, we learned its correlated words ($w_{\bar{s}}$), where $\bar{s}$ is a script different from s. We opted for a continuous bag of words (CBOW) method [12]. For each w_s, we keep its n best correlated words $w_{\bar{s}}$. Then CWL has been exploited in the matching process of documents to produce comparable comments. This process has been iterated to improve, at each step, the quality of the supposed comparable documents (Figure 3). This method achieved good results and allowed us to build a comparable Algerian dialect corpus named CALYOU.

In Figure 4, we plot the result of the comparability in terms of Recall in order to retrieve the best value of the hidden layer (N) of the CBOW algorithm. The curves show that the best value of N is 200 obtained at iteration 2 of the Word2Vec process. Another important parameter is the window-size (ws), which has been determined by making several experiments and by fixing N to 200. The best parameter, in accordance to Figure 5 is equal to 100. These parameters have been tested on a tuning corpus of 310 comparable comments.

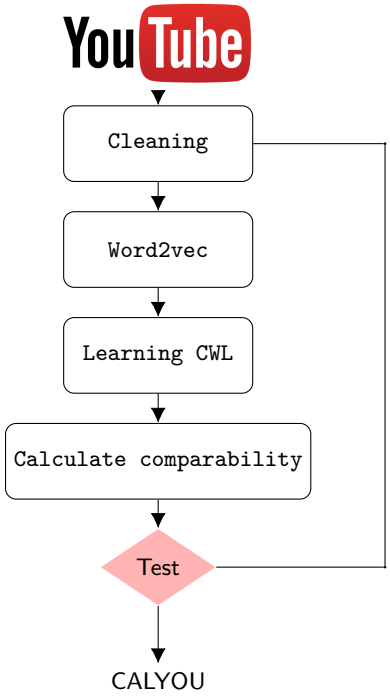

Figure 3: Iterative word embeddings training algorithm

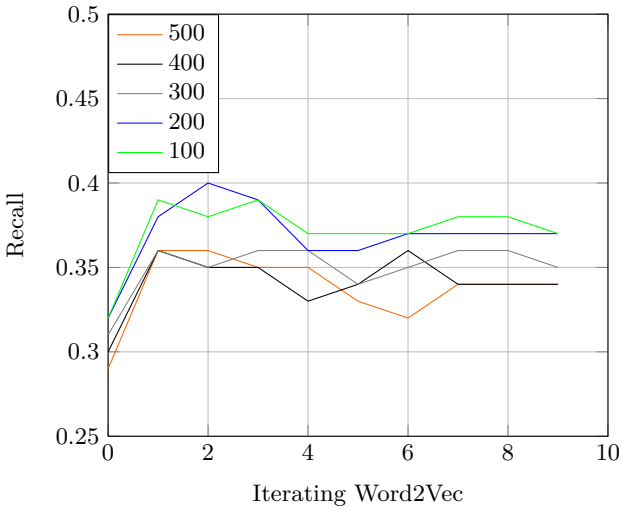

Figure 4: Evolution of the Recall in accordance to the number (N) of neurons in the hidden layer and in terms of Word2Vec iterating process

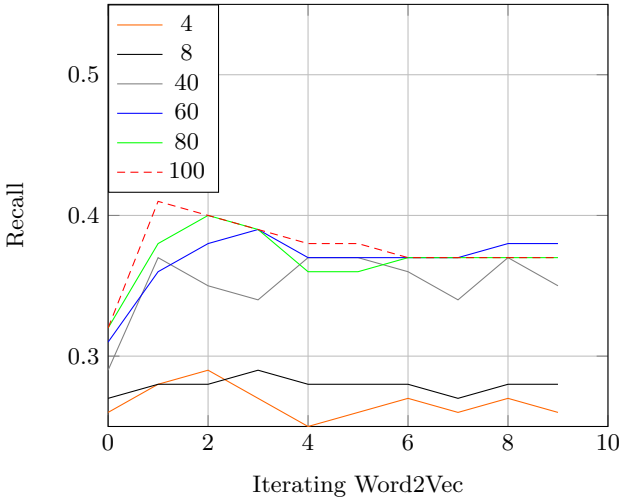

Figure 5: Evolution of the Recall in accordance to the window-size parameter and in terms of Word2Vec iterating process

5. Investigating CALYOU

Such as in Section 3.1, we present in the following some figures concerning the learned comparable corpora CALYOU. The result of CALYOU consists in a list of pairs of comments, for which the source is written in Latin script (Arabizi, French or English) and the target is written in Arabic (MSA or dialect). An example, extracted from CALYOU is given in Table 5.

Table 6 shows that the repartition of Arabizi in the source side of CALYOU is very high (97.3%), while the percentage of comments in French constitutes 2.59% of comments. Concerning the target side, 81% of comments are in Arabic dialect and 19% are in MSA. In the pie chart of Figure 6, such as in the Youtube corpus (Figure 1), the highest frequency distribution concerns the Algerian dialect written in LS. The second highest frequency distribution of words in CALYOU concerns entries written in MSA (21%).

Source j'ai trop aimé la tenu c mon style **Translation** I like too much your outfit, this is my style **Target** عجبوني بزاف نحب هاد ستيل **Translation** I like them too much, I like this style
Source Deradji reste avec le sport **Translation** Deradji continue taking care of sport **Target**: يا سي دراجي انت مختص في الرياضة ابقا في الرياضة نبقو نحبوك **Translation** Mr derradji you are expert in Sport, please continue in sport and we will continue appreciating you
Source radja meziane vraiment cette chanson djat thebel be la voix dialek w rabi yerhem kamel messaoudi **Translation** Raja Meziane this song is wonderful with your voice, may God bless the soul of Kamel Messaoudi **Target** كلمات روعة وجات مع صوتك ربي يرحمو كمال مسعودي **Translation** Beautiful Lyrics especially with your voice, may God bless the soul of Kamel Messaoudi

Table 5: Example of comparable comments extracted from CALYOU

	Arabizi	MSA	AD	FR	EN
Source (%)	97.3	-	-	2.59	0.04
Target (%)	-	19	81	-	-

Table 6: Figures on comparable comments of CALYOU

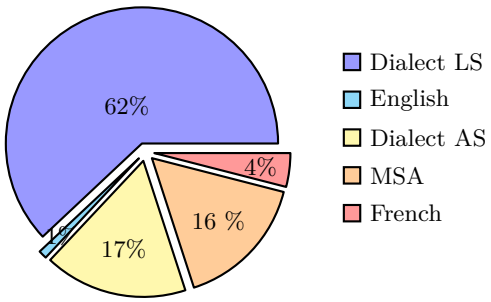

Figure 6: CALYOU Vocabulary

6. Conclusion

In this paper, we presented an analysis of the corpus collected from Youtube (more than 17M of words). The study of this corpus shows that Algerian people, in this social network, use the Latin script intensively. In fact, 47% of the comments are written with this script that was a real surprise for us. To reinforce this observation, we noticed that the percentage of distinct dialect words written in Arabizi is also high (46%) excluding those in French and English. Another crucial issue is the strong presence of the code-switching in the corpus. In fact 82% of the comments are a mixture of several varieties of languages, while only 9% of the comments are entirely written in dialect.

We proposed a multilingual word embedding approach to extract from the latter corpus a comparable one (CALYOU). In other words, each comment in Latin script is aligned with the best corresponding one in Arabic script. We trained, on a tuning corpus the parameters of the

CBOW method, then with the best parameters, we got a performance of 41% in terms of Recall by using iteratively the Word2Vec approach.
Finally, CALYOU is composed by 325K pairs of comparable comments, 62% of its vocabulary is in Arabizi and only 17% is written in Arabic script.

7. References

[1] K. Abidi, M. A. Menacer, and K. S. and, "Calyou: A comparable spoken algerian corpus harvested from youtube," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association,Sweden August 20-24 2017, 2016*, 2017.

[2] R. Zbib, E. Malchiodi, J. Devlin, D. Stallard, S. Matsoukas, R. M. Schwartz, J. Makhoul, O. Zaidan, and C. Callison-Burch, "Machine translation of arabic dialects," in *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada*, 2012, pp. 49–59.

[3] K. Almeman, M. Lee, and A. A. Almiman, "Multi dialect arabic speech parallel corpora," in *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, 2013.

[4] H. Mubarak and K. Darwish, "Using twitter to collect a multi-dialectal corpus of arabic," in *The EMNLP 2014Workshop on Arabic Natural Language Processing 1?7*, 2014.

[5] H. Bouamor, N. Habash, and K. Oflazer, "A multidialectal parallel corpus of arabic," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, 2014, pp. 1240–1245.

[6] K. Meftouh, S. Harrat, S. Jamoussi, M. Abbas, and K. Smaïli, "Machine translation experiments on PADIC: A parallel arabic dialect corpus," in *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, PACLIC 29, Shanghai, China, October 30 - November 1, 2015*, 2015.

[7] S. Harrat, K. Meftouh, M. Abbas, and K. Smaïli, "Building Resources for Algerian Arabic Dialects," in *15th Annual Conference of the International Communication Association Interspeech.* Singapour, Singapore: ISCA, Sep. 2014.

[8] S. Harrat, K. Meftouh, M. Abbas, S. Jamoussi, M. Saad, and K. Smaili, "Cross-Dialectal Arabic Processing," in *International Conference on Intelligent Text Processing and Computational Linguistics*, ser. Lecture Notes in Computer Science, cairo, Egypt, Apr. 2015.

[9] A. Douib, D. Langlois, and K. Smaili, "Genetic-based decoder for statistical machine translation," in *Springer LNCS series, Lecture Notes in Computer Science*, Dec. 2016.

[10] M. Menacer, O. Mella, D. Fohr, D. Jouvet, D. Langlois, and K. Smaili, "Development of the arabic loria automatic speech recognition system (alasr) and its evaluation for algerian dialect," in *Third International Conference On Arabic Computational Linguistics, Dubai, November 2017*, 2017.

[11] A. Dey and P. Fung, "A hindi-english code-switching corpus," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, 2014, pp. 2410–2413.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.